Cancer Epidemiology 36 (2012) 425–429

Contents lists available at SciVerse ScienceDirect

Cancer Epidemiology

The International Journal of Cancer Epidemiology, Detection, and Prevention

journal homepage: www.cancerepidemiology.net

Validity of cancer diagnosis in a primary care database compared with linked cancer registrations in England. Population-based cohort study

A. Dregan [a,*], H. Moller [b], T. Murray-Thomas [c], M.C. Gulliford [a]

[a] King's College London, Department of Primary Care and Public Health, 42 Weston Street, London SE1 3QD, United Kingdom
[b] King's College London, Thames Cancer Registry, 42 Weston Street, London SE1 3QD, United Kingdom
[c] The General Practice Research Database Group, MHRA, 151 Buckingham Palace Road, London SW1W 9SZ, United Kingdom

ARTICLE INFO

Article history:
Received 23 April 2012
Received in revised form 28 May 2012
Accepted 29 May 2012
Available online 21 June 2012

Keywords:
Cancer
Epidemiologic methods
Registries
Electronic health records
Primary care

ABSTRACT

Aims: The present study aimed to evaluate the validity of cancer diagnoses and death recording in a primary care database compared with cancer registry (CR) data in England.
Methods: The eligible cohort comprised 42,556 participants, registered with English general practices in the General Practice Research Database (GPRD) that consented to CR linkage. CR and primary care records were compared for cancer diagnosis, date of cancer diagnosis and death. Read and ICD cancer code sets were reviewed and agreed by two authors.
Results: There were 5216 (91% of CR total) cancer events diagnosed in both sources. There were 494 (9%) diagnosed in CR only and 213 (4%) that were diagnosed in GPRD only. The predictive value of a GPRD cancer diagnosis was 96% for lung cancer, 92% for urinary tract cancer, 96% for gastro-oesophageal cancer and 98% for colorectal cancer. 'False negative' primary care records were sometimes accounted for by registration end dates being shortly before cancer diagnosis dates. The date of cancer diagnosis was median 11 (interquartile range −6 to 30) days later in GPRD compared with CR. Death records were consistent for the two sources for 3337/3397 (99%) of cases.
Conclusion: Recording of cancer diagnosis and mortality in primary care electronic records is generally consistent with CR in England. Linkage studies must pay careful attention to selection of codes to define eligibility and timing of diagnoses in relation to beginning and end of record.

1. Introduction

Electronic health records are an increasingly utilised resource for epidemiological research. In the UK, records from primary care databases have been used in studies of cancer diagnosis and prognosis [1]. Validation studies have confirmed the accuracy and completeness of UK electronic patient records with respect to several clinical conditions as well as pharmacological treatment, and death [2–9]. In the UK, a national system for cancer registration aims to record all new cancer diagnoses. Cancer Registry (CR) data are considered to represent an accurate resource for studies of cancer incidence and prognosis [10]. The validity of cancer diagnoses in primary care electronic records in comparison with cancer registrations has not been well described.

The present study builds on an earlier analysis of data from the General Practice Research Database (GPRD) [1] that evaluated the incidence of cancer in patients presenting with four 'alarm' symptoms, haematuria, haemoptysis, dyphagia and rectal bleeding. At the time of the initial analysis linked cancer data were not available to ascertain the frequency and validity of cancer diagnoses recorded in the GPRD. In order to confirm our initial findings, we wanted to ascertain the validity of cancer diagnoses and dates of cancer diagnosis in GPRD. We have made use of the opportunity presented by a novel linkage between cancer registrations with primary care electronic records to compare data from the two sources. The present report therefore aims to evaluate the validity of cancer diagnoses in primary care electronic health records by comparing the occurrence, and timing of cancer diagnoses between GPRD with cancer registrations. We specifically evaluated diagnoses of lung cancer, colorectal cancer, cancer of the oesophagus or stomach and urinary tract cancers.

2. Methods

2.1. Data

2.1.1. Cancer registry

Cancer registries in England represent the only available source of reliable population-based data on cancer incidence, prevalence

* Corresponding author at: Department of Primary Care and Public Health, 6th Floor Capital House, 42 Weston Street, King's College London, London SE1 3QD, United Kingdom. Tel.: +44 0207 8486639; fax: +44 0207 8486620.
E-mail address: alexandru.dregan@kcl.ac.uk (A. Dregan).

1877-7821/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.canep.2012.05.013

and survival, excluding non-melanoma skin cancer which is not collected systematically. Information is collected on new diagnoses of cancer from hospitals, pathology laboratories, hospices, cancer treatment centres, cancer screening programmes, Hospital Episode Statistics (HES), cancer waiting times (CWT) and death certificates. Within hospitals, data are collected from several sources (e.g. pathology departments, medical records, and radiotherapy databases) increasingly using electronic data sources. Data are checked for consistency and quality assured.

2.1.2. General Practice Research Database

The GPRD is a large computerised database of anonymised longitudinal medical records from primary care. Currently data are being collected from more than 600 general practices on about 5 million active patients of research standard. At the time of this study the GPRD covered approximately 7% of the UK population.

2.1.3. Data linkage

The GPRD data linkage scheme is currently confined to England and cancer data are only available for the subset of English practices which have consented to participate in the scheme. Cancer registry (CR) records were linked to GPRD data by a Trusted Third Party (TTP) using a deterministic algorithm based on the patient National Health Service (NHS) number, post code, gender and date of birth information. The variables used for data linkage were available in both data sources. Data on all cancer diagnoses as these exist in the CR were obtained from the merged cancer registry dataset which contains data from all eight regional registries in England, complete to the end of 2006. Death data as reported in the cancer registry was obtained from the linked Office of National Statistic (ONS) minimum dataset which is supported by the cancer registry.

2.2. Study population

This study was based on the subset of English practices in the wider GPRD study that were eligible to be linked to Cancer registry data. Only those practices which continuously contributed data to the GPRD during the period 1st January 2001–31st December 2007 were eligible for inclusion. Individual participants were included if they had at least 12 months of up-to-standard follow-up (UTS; practices with records considered to be of research standard with respect to the completeness of recording and diagnosis accuracy) prior to the start of observation on 01/01/2002 and were further eligible if they had no alarm symptom or cancer diagnosis documented on or before 31st December 2001 in their clinical or referral record. Participants were included in the sample if they had recorded symptoms of haematuria, haemoptysis, dysphagia or rectal bleeding in GPRD or were diagnosed with lung cancer, colorectal cancer, gastro-oesophageal cancer or cancer of the urinary tract in either GPRD or CR. The study sample was a convenience sample from a study cohort that will consider cancer risk in relation to alarm symptoms. The presence of alarm symptoms was used to define a sample with a high frequency of cancer, but the presence or absence of alarm symptoms was not considered further in this analysis. The codes for alarm symptoms are detailed in a previous publication [1]. The latest data collection date in the GPRD sample was up to May 2008. Cancer registrations were obtained in November 2010 from a dataset that included cancer registrations up to the end of 2006. For the purposes of the present study, data were analysed for cancer diagnoses recorded between 01/01/2002 and 31/12/2006.

2.2.1. Adjudication of cancer diagnoses in the GPRD and cancer registry

Two separate lists, one containing Read/Oxford Medical Information System (OXMIS) codes suggestive of cancer as documented in the GPRD and one comprising of CR International Statistical Classification of Diseases and Related Health Problems 10th revision (ICD-10) [11] and the International Classification of Diseases for Oncology (ICD-O-2) codes were reviewed by two reviewers (AD and HM) with experience in primary care epidemiology and cancer registration respectively. Raters reviewed each code and determined whether use of the code was indicative of a relevant cancer diagnosis. The selection criteria used to identify codes of interest to this study included whether the clinical term associated with a code made reference to the cancer site and whether terms gave some indication of malignancy. Codes a priory judged to be inconclusive were excluded from subsequent analyses (Appendix 1).

2.3. Statistical analysis

For each participant in the sample, the analysis determined whether cancer was recorded or not in the GPRD and CR data respectively, for each of the four cancer groups included in the study: lung, colorectal, gastro-oesophageal and urinary tract. The positive predictive value (PPV; proportion of GPRD cases confirmed in CR), sensitivity (proportion of cases correctly identified in GPRD), and specificity (proportion of cases correctly identified as cancer-free in GPRD) were estimated using CR as the reference data. Median and interquartile ranges for the difference in date of cancer diagnosis between GPRD and CR databases were estimated for the four cancer groups. The validity of death and the index date of death recording in the GPRD using CR as the reference data was also performed using the same procedures. Because the available CR data included only month and year of cancer diagnosis, a day of diagnosis for each CR case was imputed.

3. Results

The initial GPRD sample comprised 83,841 participants. Fig. 1 provides a flowchart of the sample selection process. There were 173 (53%) out of the original 334 GPRD practices without any cancer occurrence recorded in the CR that were excluded from further analyses as not consenting to linkage. This left 158 (47%) eligible practices that participated in linkage between GPRD and CR. Consequently, 49% ($N = 37,283$) of participants with alarm symptoms ($N = 76,143$) and 46% ($N = 5254$) participants with cancer diagnoses ($N = 11,351$) from the wider GPRD study were excluded from the analysis. The reviewed Read and OXMIS code list led to the further exclusion of 151 cancer patients as several of the Read and OXMIS codes used in the original study were judged not to indicate a malignant neoplasm (Appendix 1).

In the linked CR data, out of an initial sample size of 8620 cancers, 2249 (27%) cancer cases were excluded because the site of cancer recorded was unrelated to the study objective (e.g. breast, prostate, skin, brain), 116 (1%) cancer cases were excluded because they were outside the study period (i.e. diagnosis made prior to 2002 in the CR) and 82 (1%) cancer cases were excluded because the site was judged not to indicate a relevant malignant neoplasm (Appendix 1). A further 525 (8%) CR records were excluded because the cancer type and date of cancer diagnosis were missing. Matched records for the 525 cases with missing CR data were also excluded from the GPRD dataset. Application of these eligibility criteria resulted in a final study sample of 42,556 participants who were recorded with alarm symptoms in GPRD and/or with related cancer diagnoses in either CR or GPRD during the study period.

There were 5429 cancer diagnoses in GPRD and 5710 in the CR, with 5216 (91% of CR total) diagnosed in both sources. There were 494 (9%) diagnosed in CR but not in GPRD and 213 (4%) that were diagnosed in GPRD but not CR. Urinary tract cancers accounted for 78/213 (37%) of the GPRD-only cancers. The positive predictive